Czech Semi-Automatic Dictionary

Marek Blahuš, Michal Cukr, Miloš Jakubíček, Vojtěch Kovář, František Kovařík

8–10th Dec 2023

Semi-automatic dictionary making

1. Automatic data acquisition.
2. Manual annotation and revision.

Dictionary Express projects

- From corpora.
- Previous projects:

  Urdu, Lao, Tagalog, Ukrainian.
- Headwords, pronunciation, translation, thesaurus, examples.

Dictionary Express

- Main feature: **rapid** creation.
- Content checked by editors (annotators)
- Project management by coordinators (supervisors).

Coordinator vs. editor (annotator)

Editors = native speakers.

X

Coordinators **needn’t** know the language.

Aim of Czech Dictionary Express (CDE)

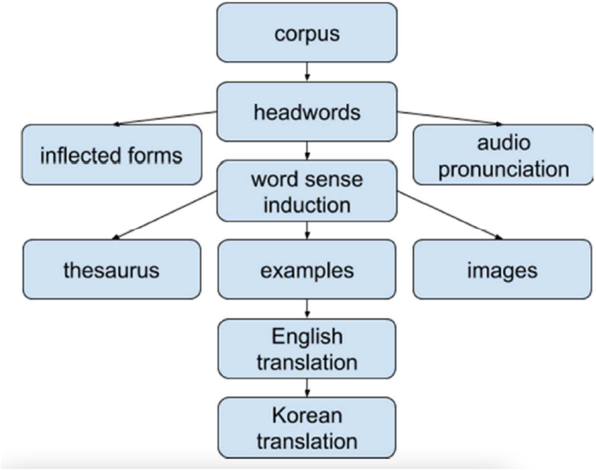
- A new Czech dictionary (obviously).

Aim of Czech Dictionary Express (CDE)

- A new Czech dictionary (obviously).
- Examining the **methodology**.

  Coordinators **are** native speakers.

Project phases



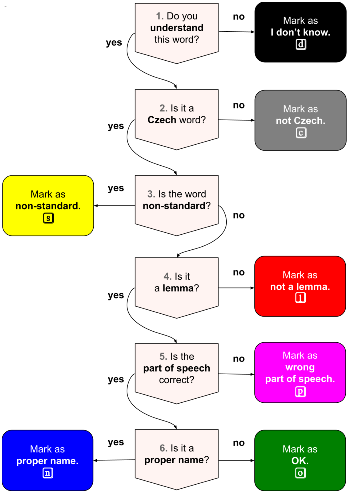
Project preparation

- Automatic: Generate headword **batches**.

  1 batch = 1000 headwords.

- Annotator recruitment.

Headword annotation



Language-related problems 1/2

- Only single words (*bát*-verb).
  No context.

- Presumption of correctness (*prostřednictvím*-preposition, *hajný*-adjective).

- Abbreviations (*dr.*-noun, *např.*-adverb, *cca*-adverb).

Language-related problems 2/2

- (Non-)negated lemmas
  (*nenávidět*, *odmyslitelně*).

- Interjections
  (*kikirikí*/*kykyryký* vs. *kykyrykýhyhý*).

Extending the lexicon

- 15 000 –> 80 000 headwords,

  2x annotated.

- Fast & furious.

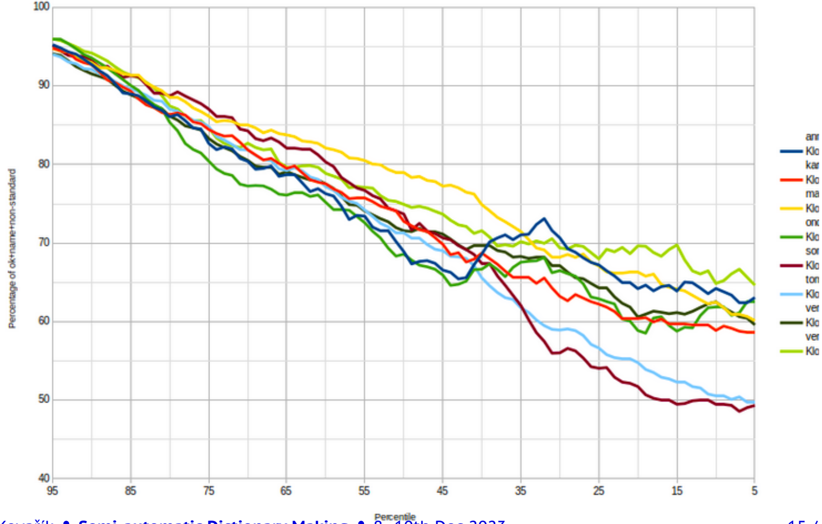
  Bigger corpora. Better corpus tools.
  => More OK headwords.

(a) **Inter-annotator agreements**

1/1:	8999
1/2:	13340
2/2:	47695
1/3:	85
2/3:	539
3/3:	8342
3/8:	14
4/8:	57
5/8:	102
6/8:	112
7/8:	194
8/8:	521

(b) **Flag statistics (more than 50 % agreement)**	
OK	43012
name	4923
not a lemma	3544
non-standard	336
wrong POS	583
I don’t understand	1947
not Czech	3160

Correctness – frequency



Percentage of ok+name+non-standard per percentile (80K headwords by docf) – running average (10 percentiles)

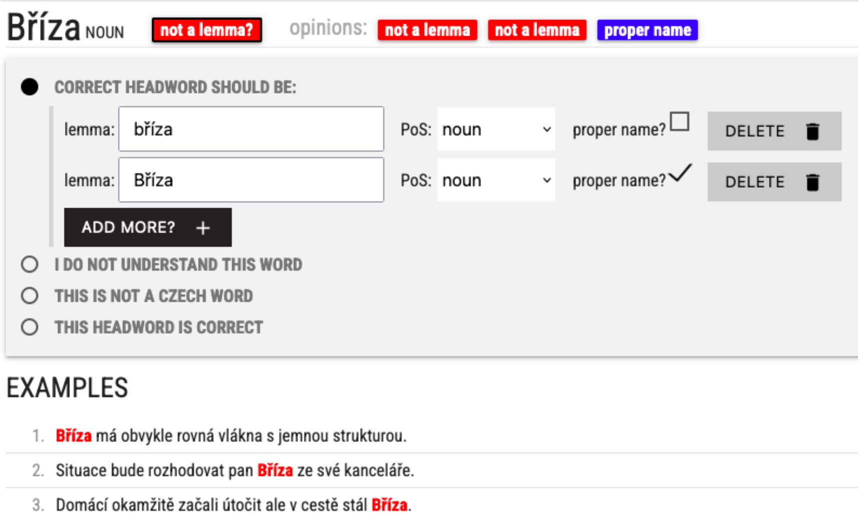
Difficulties 1/2

- Diminutives (*lesík*), gender (*stolařka*).
- Views of the standard (*tož*, *-ismus/-izmus*).
- Deciding POS (*zima*-adverb, *výborná*-noun).

Difficulties 2/2

- Existence of the non-negated form (*překonatelný*, *zbytný*.).
- Czech vs. foreign word (*link*, *Trump*).

Next phase: Headword revision



Bříza NOUN not a lemma? opinions: not a lemma not a lemma proper name

● CORRECT HEADWORD SHOULD BE:

lemma: bříza PoS: noun proper name? □ DELETE

lemma: Bříza PoS: noun proper name? ✓ DELETE

ADD MORE? +

○ I DO NOT UNDERSTAND THIS WORD

○ THIS IS NOT A CZECH WORD

○ THIS HEADWORD IS CORRECT

EXAMPLES

1. **Bříza** má obvykle rovná vlákna s jemnou strukturou.
2. Situace bude rozhodovat pan **Bříza** ze své kanceláře.
3. Domácí okamžitě začali útočit ale v cestě stál **Bříza**.

Nexter phase: Word forms

hajný (noun) + context (optional)

- hajného – **correct?**
- hajnému – **correct?**
- hajnímu – **correct?**

Conclusion

We are examining the methodology...

... while creating a rapid Czech dictionary...

... with 80 000 annotated headwords.

Thanks for your attention.

Do you have any **questions**?

MASARYKOVA UNIVERZITA